Michał Bernardelli[‡]

Ocena korelacji wyników testu koniunktury z wykorzystaniem ukrytych modeli Markowa

Streszczenie

W artykule opisano zastosowanie nowej miary zależności opartej na ukrytych modelach Markowa oraz ścieżkach Viterbiego, do badania stopnia korelacji pomiędzy szeregami sald odpowiedzi respondentów na pytania testu koniunktury w przemyśle prowadzonego przez Instytut Rozwoju Gospodarczego Szkoły Głównej Handlowej w Warszawie. Wyznaczone wartości współczynników nowej korelacji zestawiono z klasyczną korelacją Pearsona. Dokonano porównania na przykładowych parach szeregów, będącego empirycznym dowodem przydatności nowej miary na danych ankietowych. W wielu przypadkach proponowany pomiar podobieństwa między szeregami okazał się bardziej adekwatny. Co więcej, zastosowanie bardziej wyrafinowanej metody pozwala na identyfikację okresów podobieństwa oraz okresu większego zróżnicowania analizowanych szeregów.

Słowa kluczowe: analiza szeregów czasowych, miary podobieństwa, ukryte modele Markowa, ścieżki Viterbiego, zależność statystyczna

Kod klasyfikacji JEL: C22, E32

[‡] Instytut Ekonometrii, Szkoła Główna Handlowa w Warszawie

Correlation analysis of survey data with the use of hidden Markov models

Abstract

The paper proposes a new measure of the similarity between time series, based on hidden Markov models and Viterbi paths. The results are compared with the Pearson correlation coefficient. The comparison shows that the proposed measure gives more accurate estimates of the similarity and has some advantages over other measures commonly used, namely, it identifies periods (subsamples) of high and low similarity between time series.

Keywords: hidden Markov models, similarity measures, statistical relationship, time series analysis, Viterbi paths

JEL classification: C22, E32

1. Wprowadzenie

Istnieją różne metody badania zależności między zmiennymi. Najpopularniejsze z nich polegają na obliczeniu współczynnika takiej korelacji, która odpowiada rodzajowi badanych zmiennych (zmienne jakościowe, ilościowe ciągłe, ilościowe dyskretne). Część z metod jest znana od ponad kilkudziesięciu lat (Soper i in., 1917, Kendal & Stuart, 1973), przy czym najczęściej stosowaną miarą jest współczynnik korelacji liniowej Pearsona, zaproponowany w 1895 roku przez Francisa Galtona i Karla Pearsona. Ta miara ma pewne znane wady, m.in. służy ocenie tylko liniowych zależności, a w dodatku jest wrażliwa na obserwacje odstające. Zaproponowano więc wiele innych metod porównywania zmiennych (zob. Székely i in., 2007, Tjostheim & Hufthammer, 2013). Jedna z bardziej zaawansowanych obliczeniowo metod (Bernardelli, 2018) wykorzystuje koncepcję ukrytych modeli Markowa (HMM) oraz ścieżek Viterbiego do określenia stopnia zależności pomiędzy szeregami czasowymi. Zaproponowana metoda w wielu przypadkach wydaje się – poprzez identyfikację okresów zbieżności i rozbieżności – odzwierciedlać faktyczne podobieństwo między szeregami czasowymi.

Celem artykułu było sprawdzenie stopnia zależności odpowiedzi ankietowanych na poszczególne pytania testu koniunktury w przemyśle przetwórczym, prowadzonego przez Instytut Rozwoju Gospodarczego Szkoły Głównej Handlowej w Warszawie. Do badania zależności wykorzystano miarę opartą na HMM oraz ścieżkach Viterbiego. Wyniki zestawiono ze współczynnikami korelacji Pearsona. Badanie to może dać rozwiązanie dwóch kwestii. Po pierwsze, jest to weryfikacja przydatności nowej miary do analizy danych ankietowych. Proponowana miara nie jest bowiem uniwersalna i została w zamierzeniu skonstruowana dla danych o charakterze makroekonomicznym. Po drugie, wyniki badania stanowią dodatkowy test przydatności pytań pod względem braku redundantności. Gdyby bowiem odpowiedzi na dwa różne pytania niosły te same informacje o zmienności w ocenach respondentów, to sensowność testu koniunktury w przemyśle w takiej postaci byłaby wątpliwa.

Artykuł składa się z sześciu części. Po wprowadzeniu, w rozdziale drugim została przedstawiona teoria ukrytych modeli Markowa oraz opis ścieżek Viterbiego. Matematyczne sformułowania zostały przy tym ograniczone na rzecz przedstawienia idei stosowalności oraz przykładów zastosowań. W kolejnym, trzecim rozdziale, zawarto opis miary zależności opartej na HMM oraz ścieżkach Viterbiego. Na przykładzie przedstawiono porównanie nowej miary ze współczynnikiem korelacji Pearsona. Krótki opis testu koniunktury w przemyśle IRG oraz pytań ankietowych wchodzących

w jego skład stanowi zawartość rozdziału 4, zaś wyniki obliczeń (wartości współczynników korelacji Pearsona oraz HMM) wraz z przykładowymi wykresami znajdują się w rozdziale 5. Artykuł kończy się podsumowaniem.

2. Ukryte modele Markowa i ścieżki Viterbiego

W pracy wykorzystane zostały ukryte modele Markowa (*hidden Markov models*, HMM), znane też pod nazwą przełącznikowych modeli Markowa (Cappé i in., 2005). Formalna definicja określa dwa warunki, które musi spełniać częściowo obserwowalny proces $\{(X_t, Y_t)\}_{t=1}^{\infty}$:

1. Składowa nieobserwowalna $\{X_t\}_{t=1}^{\infty}$ jest jednorodnym łańcuchem Markowa ze skończoną przestrzenią stanów S.
2. Obserwowalne zmienne losowe Y_1, Y_2, …, Y_t są pod warunkiem $(X_1, X_2, \ldots, X_t)$ niezależne, przy czym rozkład zmiennej losowej Y_t pod tym warunkiem zależy jedynie od zmiennej losowej X_t.

HMM są od lat wykorzystywane w rozpoznawaniu pisma czy mowy, a z nowszych zastosowań można wskazać sekwencjonowanie DNA. W dziedzinie ekonomii, ukryte modele Markowa stosowane są jako narzędzie analizy szeregów czasowych, jak również w badaniach koniunktury, np. do identyfikacji punktów zwrotnych czy badania synchronizacji cykli koniunkturalnych. HMM mogą mieć jednak zastosowanie wszędzie tam, gdzie na podstawie jakichś sygnałów (obserwowalnego szeregu czasowego) chcemy wyznaczyć ukryty wzorzec (łańcuch Markowa).

W niniejszym badaniu zostały zastosowane dwustanowe modele z jednowymiarowymi normalnymi rozkładami warunkowymi, to jest $S = \{0,1\}$ oraz

$$Y_t|_{X_t=0} \sim N(\mu_0, \sigma_0), \quad Y_t|_{X_t=1} \sim N(\mu_1, \sigma_1),$$

gdzie $\mu_0 < \mu_1$. Interpretacja stanów jest zależna od charakteru analizowanego szeregu, ale można przyjąć, że w przypadku badań koniunktury stan 0 odpowiada okresowi dekoniunktury, zaś stan 1 poprawie sytuacji. Możliwe są badanie większej liczby stanów (Bernardelli, 2014) lub wielowymiarowych rozkładów warunkowych (Bernardelli & Dędys, 2017), jednak proponowana miara zależności, opisana w kolejnym rozdziale, korzysta z modeli o dwóch stanach, więc w opisie ograniczono się tylko do takich modeli.

Parametry HMM można obliczyć wykorzystując iteracyjny algorytm Bauma-Welcha, który mimo deterministycznego charakteru, może dawać wyniki dalekie od optymalnego. Wyniki te zależą bowiem od przyjętych początkowych wartości prawdopodobieństw. W celu zwiększenia szans na

znalezienie globalnego optimum standardowo wykonuje się wielokrotnie obliczenia dla tych samych danych, ale różnych wartości startowych. Na temat kryteriów wyboru najlepszego modelu oraz opisu parametrów modelu można przeczytać np. w (Bernardelli, 2014, Bernardelli & Dędys, 2014). W badaniu, przedstawionym w tej pracy liczba symulacji była równa 2000, ze względu na stabilność obliczeń numerycznych oraz niewielką liczbę stanów.

Wyznaczenie wartości parametrów modelu jest tylko pierwszym z dwóch zadań niezbędnych do znalezienia ciągu ukrytych stanów. W wyniku użycia algorytmu Bauma-Welcha otrzymujemy zestaw prawdopodobieństw, na podstawie których należy podjąć decyzję co do konkretnej ścieżki stanów. Istnieje kilka algorytmów umożliwiających określenie takiej ścieżki, ale z punktu widzenia ekonomicznej interpretacji najbardziej odpowiedni wydaje się algorytm Viterbiego, który wyznacza najbardziej prawdopodobną, przy danym sygnale, ścieżkę przebytą przez ukryty łańcucha Markowa w całym rozpatrywanym okresie. Ścieżka ta nazywana jest ścieżką Viterbiego.

Połączenie algorytmów Bauma-Welcha oraz Viterbiego wyznacza ścieżkę stanów złożoną z 0 i 1, która odpowiada szeregowi czasowemu, zaś chwile zmiany stanów mogą być interpretowane jako punkty zwrotne. Ścieżki Viterbiego są podstawą konstrukcji proponowanej miary zależności, która została zastosowana w tej pracy do badania podobieństwa odpowiedzi respondentów na pytania testu koniunktury w przemyśle. Opis tej miary zostanie przedstawiony w następnym rozdziale.

3. Współczynnik korelacji r_{HMM}

W rozdziale 2 opisana została koncepcja ukrytych modeli Markowa oraz ścieżek Viterbiego. Została ona wykorzystana (Bernardelli, 2018) do konstrukcji współczynnika korelacji, oznaczanego przez r_{HMM}, mającego oddawać stopień zależności pomiędzy szeregami czasowymi. Procedura obliczania r_{HMM} dla dwóch szeregów czasowych x_t i y_t długości n, może być przedstawiona w następujących krokach:

1. Normalizacja szeregów czasowych x_t i y_t

$$\widetilde{x_t} = \frac{x_t - \min\limits_{s\in\{1,\dots,n\}} x_s}{\max\limits_{\tau\in\{1,\dots,n\}} \left| x_\tau - \min\limits_{s\in\{1,\dots,n\}} x_s \right|}$$

oraz

$$\widetilde{y}_t = \frac{y_t - \min\limits_{s\in\{1,\dots,n\}} y_s}{\max\limits_{\tau\in\{1,\dots,n\}} \left| y_\tau - \min\limits_{s\in\{1,\dots,n\}} y_s \right|}$$

Po tym kroku $\widetilde{x}_t, \widetilde{y}_t \in [0;1]$.

2. Obliczenie różnicy pomiędzy znormalizowanymi szeregami czasowymi. W zależności od znaku współczynnika korelacji liniowej Pearsona *r* różnica definiowana jest

$$\widetilde{z}_t = \frac{(\widetilde{x}_t - \widetilde{y}_t) - \min\limits_{s\in\{1,\dots,n\}} (\widetilde{x}_s - \widetilde{y}_s)}{\max\limits_{\tau\in\{1,\dots,n\}} \left| (\widetilde{x}_\tau - \widetilde{y}_\tau) - \min\limits_{s\in\{1,\dots,n\}} (\widetilde{x}_s - \widetilde{y}_s) \right|}$$

dla dodatniego *r* oraz

$$\widetilde{z}_t = \frac{(\widetilde{x}_t + \widetilde{y}_t) - \min\limits_{s\in\{1,\dots,n\}} (\widetilde{x}_s + \widetilde{y}_s)}{\max\limits_{\tau\in\{1,\dots,n\}} \left| (\widetilde{x}_\tau + \widetilde{y}_\tau) - \min\limits_{s\in\{1,\dots,n\}} (\widetilde{x}_s + \widetilde{y}_s) \right|}$$

dla ujemnie skorelowanych szeregów $\widetilde{x}_t$ i $\widetilde{y}_t$.

3. Wyznaczenie ścieżki Viterbiego v_t dla szeregu $\widetilde{z}_t$.
4. Obliczenie współczynnika r_{HMM} ze wzoru

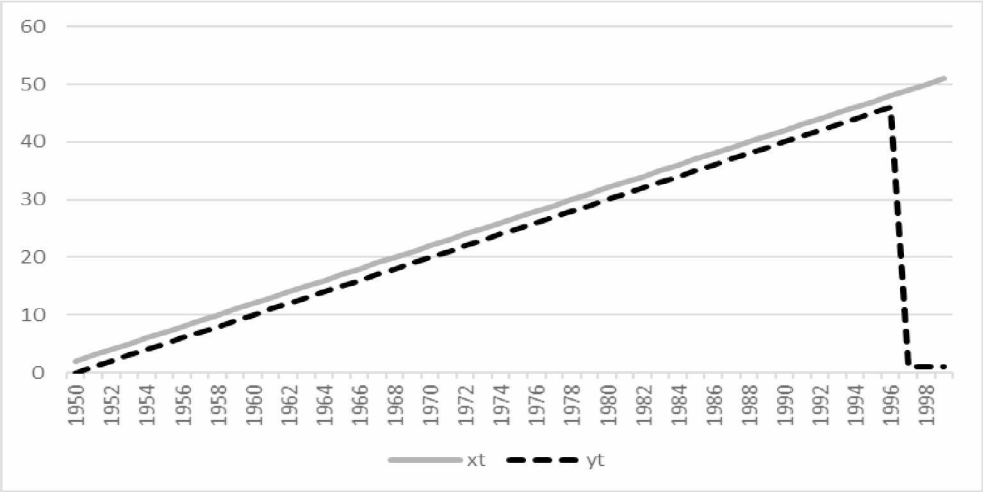
$$r_{HMM} = \frac{liczba\ stanów\ 0\ na\ ścieżce\ v_t}{długość\ ścieżki\ v_t} \in [0;1].$$

Miara r_{HMM} mówi o tym, przez jaką część badanego okresu szeregi czasowe zachowują się podobnie. Dla idealnego podobieństwa $r_{HMM} = 1$, natomiast dla szeregów, które zachowują się odmiennie przez cały badany okres $r_{HMM} = 0$.

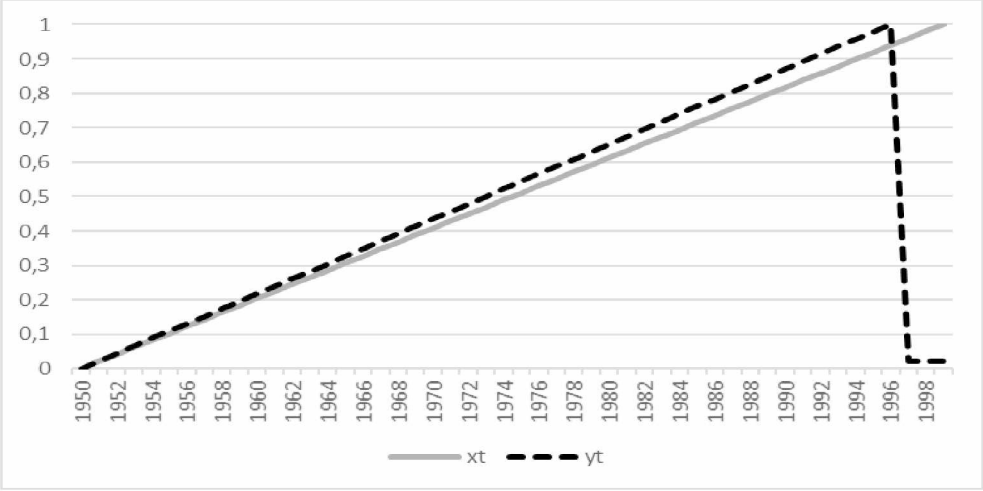
Na Rysunku 1 zostały przedstawione dwa szeregi czasowe, które są doskonale skorelowane (równoległe) przez cały okres oprócz ostatnich trzech lat. Takie dane można potraktować jako przykład odstających obserwacji, na które współczynnik korelacji Pearsona jest dość wrażliwy. Dla tych szeregów *r*=0,6951, przy czym tylko dla 3 z 50 punktów szeregi wskazują odmienne zachowanie. Proponowany współczynnik korelacji HMM osiąga inną wartość. Procedura jego obliczenia wygląda następująco.

W pierwszy kroku należy dokonać normalizacji szeregów (Rysunek 2). Są one skorelowane dodatnio (*r*=0,6951), więc do wyznaczenia szeregu $\widetilde{z}_t$

należy zastosować wzór (2). Dla tak utworzonego szeregu czasowego (krok 3) obliczana jest ścieżka Viterbiego v_t (Rysunek 3).



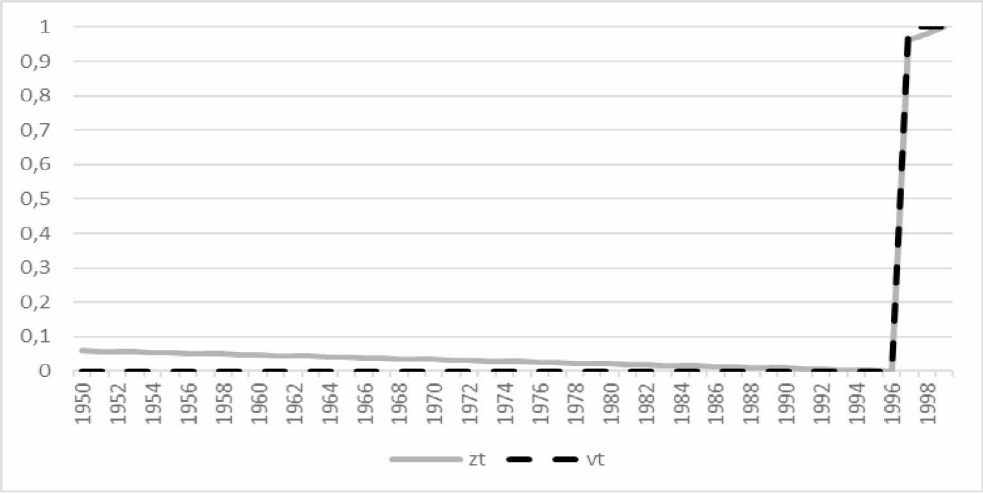
Rysunek 1. Porównywane szeregi czasowe (przykład 1).



Rysunek 2. Szeregi czasowe x_t (linia ciągła) i y_t (linia przerywana) po normalizacji (przykład 1).

Różnica między szeregami, $\tilde{z}_t$, jest przez prawie cały czas (poza ostatnimi trzema punktami) bliska zeru. Z tego powodu odpowiadająca mu ścieżka Viterbiego złożona jest ze stanów 0. Jedynie dla trzech ostatnich lat stany na ścieżce Viterbiego zmieniają się na 1. Są to lata, w których wartości szeregu $\tilde{z}_t$ znacząco rosną. Stosując wzór (4), otrzymujemy wartość $r_{HMM} =$

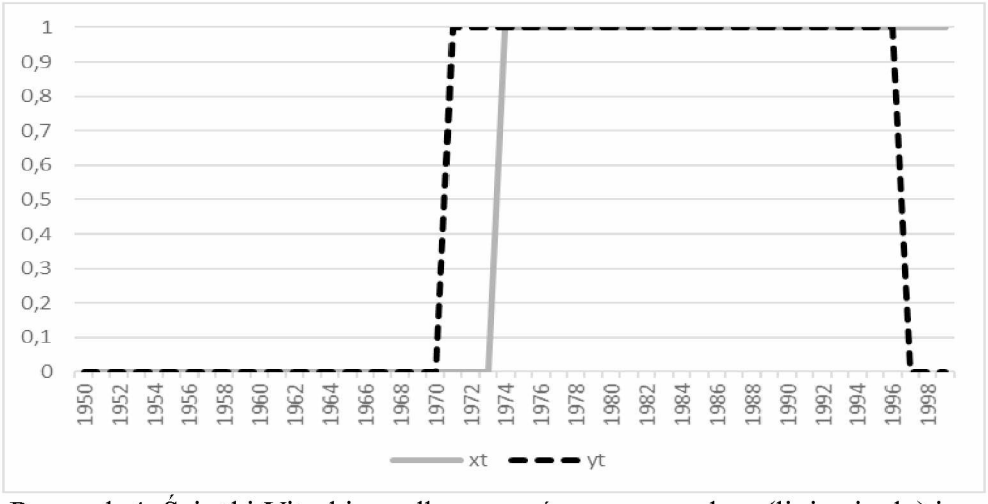
$\frac{47}{50} = 0{,}94$. W porównaniu z wartością współczynnika korelacji Pearsona równą 0,6951 – biorąc pod uwagę, iż szeregi przez dokładnie 94% okresu są równoległe – korelacja HMM daje trafniejsze przybliżenie zależności pomiędzy szeregami.



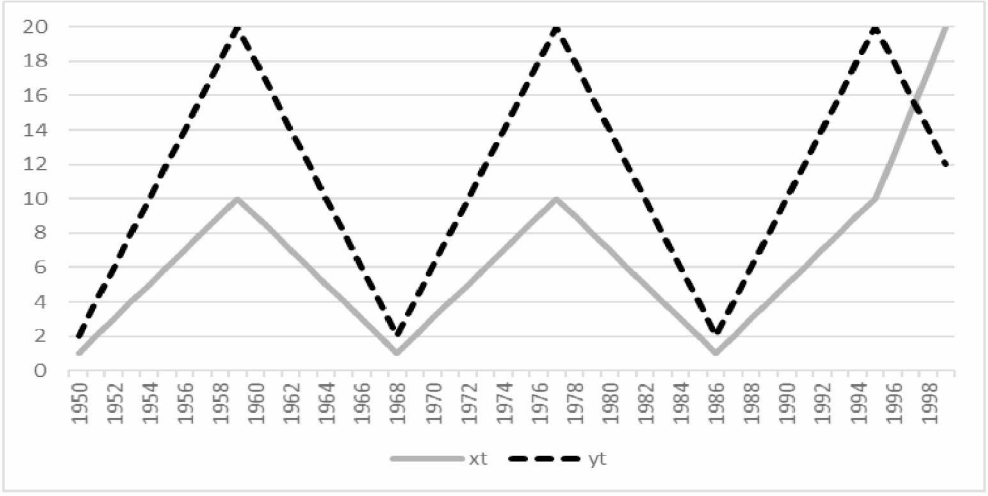
Rysunek 3. Szereg czasowy $\tilde{z}_t$ z kroku 2 (linia ciągła) oraz odpowiadająca mu ścieżka Viterbiego v_t (linia przerywana) (przykład 1).

Rozważmy alternatywną procedurę wykorzystania ukrytych modeli Markowa do konstrukcji miary $\tilde{r}_{HMM}$ podobieństwa szeregów, w której to ścieżki Viterbiego obliczane są dla każdego z porównywanych szeregów z osobna, a następnie zliczany jest procent jednakowych stanów. Jest to procedura bardziej złożona obliczeniowo, gdyż zamiast estymacji jednego modelu, wymagane jest wyznaczenie dwóch ścieżek Viterbiego, niezależnie dla każdego z szeregów. Z drugiej strony jednak zbędne są dwa pierwsze kroki zaprezentowanej w tym rozdziale procedury. Ścieżka Viterbiego dla szeregu przed i po normalizacji będzie identyczna. Nie ma też potrzeby wyznaczania różnicy szeregów. Te dwa kroki są jednak znacznie mniej czasochłonne niż krok trzeci, w którym wyznaczane są parametry ukrytego modelu Markowa. Dla danych z przykładu przedstawionego na Rysunku 1 otrzymujemy ścieżki Viterbiego przedstawione na Rysunku 4. Dokładnie 44 (z 50) stany tych ścieżek pokrywają się, co dawałoby wartość alternatywnej miary $\tilde{r}_{HMM} = \frac{44}{50} = 88\%$. Jest to wartość niższa od $\tilde{r}_{HMM} = \frac{47}{50} = 94\%$, ale wciąż trafniej oddająca rzeczywiste podobieństwo między szeregami niż współczynnik korelacji Pearsona (równy niespełna 70%).

Rozważmy jeszcze jeden przykład przemawiający na korzyść stosowania proponowanej miary r_{HMM} (zamiast alternatywnej, podobnej w założeniach miary $\tilde{r}_{HMM}$). Dwa porównywane szeregi przedstawione zostały na Rysunku 5. Fazy wzrostów i spadków są identyczne w całym przedstawianym okresie poza ostatnimi czterema latami. Dla tych dwóch szeregów czasowych wyznaczone zostaną dwie miary podobieństwa oparte na ukrytych modelach Markowa: r_{HMM} oraz $\tilde{r}_{HMM}$.

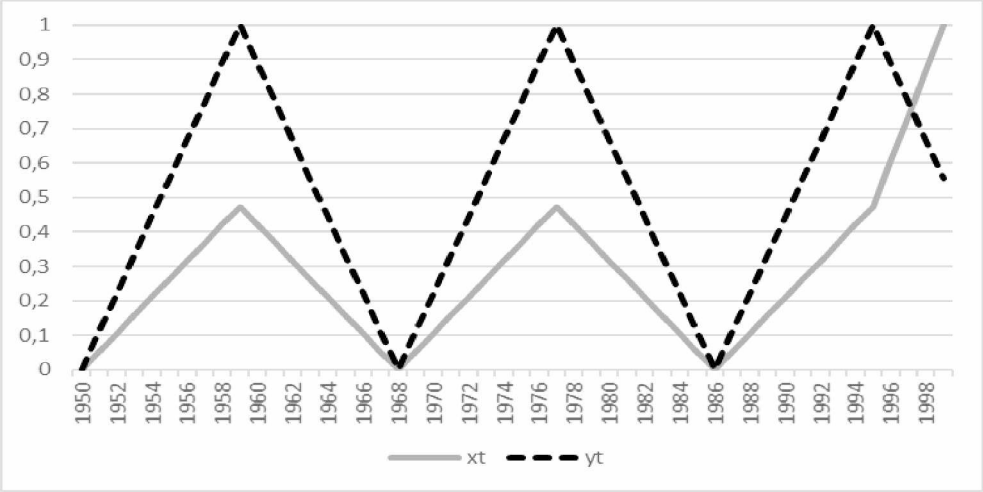


Rysunek 4. Ścieżki Viterbiego dla szeregów czasowych x_t (linia ciągła) i y_t (linia przerywana) (przykład 1).

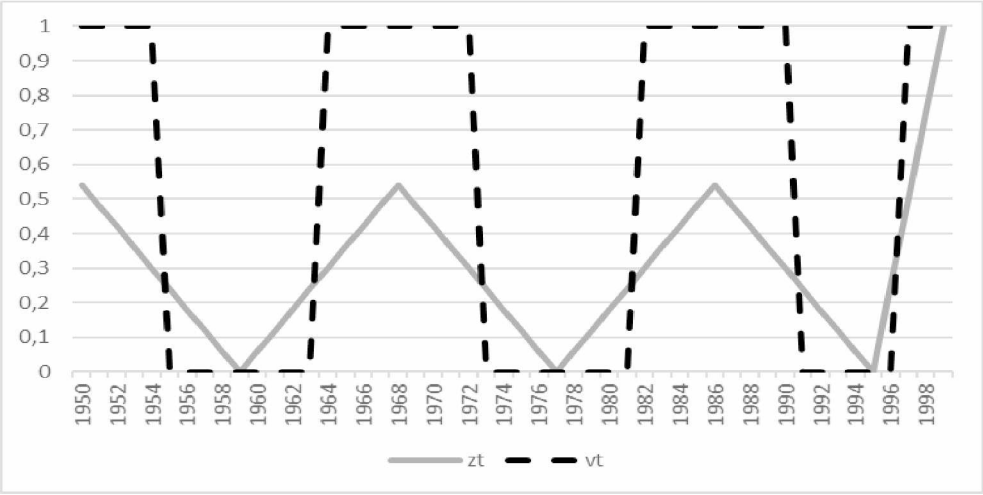


Rysunek 5. Porównywane szeregi czasowe (przykład 2).

Na Rysunku 6 przedstawione są szeregi po normalizacji. Poza skalą (oś Y), wygląd wykresów pozostaje bez zmian, a przede wszystkim niezmienne są względne różnice pomiędzy szeregami. Są one skorelowane dodatnio (r = 0,7568). Szereg czasowy różnic, wyznaczony zgodnie ze wzorem (2), oraz odpowiadająca mu ścieżka Viterbiego przedstawione zostały na Rysunku 7.



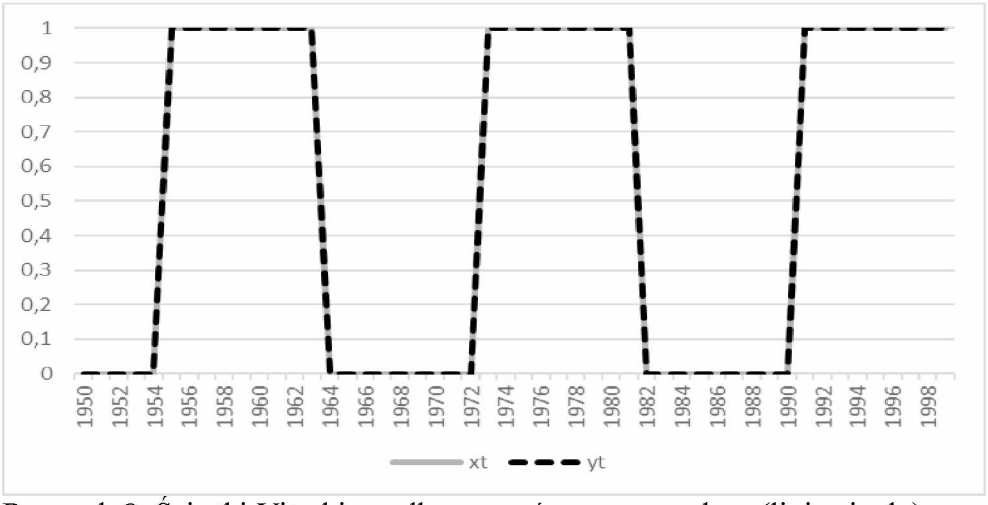
Rysunek 6. Szeregi czasowe x_t (linia ciągła) i y_t (linia przerywana) po normalizacji (przykład 2).



Rysunek 7. Szereg czasowy $\tilde{z_t}$ z kroku 2 (linia ciągła) oraz odpowiadająca mu ścieżka Viterbiego v_t (linia przerywana) (przykład 2).

Stosując wzór (4), otrzymujemy wartość $r_{HMM} = \frac{24}{50} = 0{,}48$. Jest ona znacznie niższa od wartości współczynnika korelacji Pearsona (0,7568). Przyglądając się szeregom, można jednak zauważyć, że wartości jednego z nich są równe połowie odpowiednich wartości szeregu drugiego. Jedynie w czterech ostatnich latach struktura szeregu x_t zostaje zaburzona. Szeregi te zatem znacznie różnią się od siebie, choć okresy wzrostów i spadków są prawie jednakowe.

Dokonajmy teraz porównania miary r_{HMM} z miarą alternatywną $\tilde{r}_{HMM}$ wykorzystującą koncepcję HMM. Ścieżki Viterbiego dla obu szeregów z przykładu 2 przedstawiono na Rysunku 8. Są one identyczne. Stąd z definicji miary $\tilde{r}_{HMM}$ jest ona równa 1. Oznacza to idealne dopasowanie szeregów, przy czym szeregi (patrz Rysunek 5) ewidentnie nie zachowują się jednakowo. Przykład ten ma na celu przedstawieni sytuacji, w której alternatywna miara znacznie przeszacowuje stopień podobieństwa szeregów. Mając to na uwadze, zdecydowano się na zaproponowanie pierwotnie przedstawionej miary r_{HMM}, licząc się z tym, że może ona dawać nieco zaniżone szacunki podobieństwa w stosunku do współczynnika korelacji Pearsona, jak również miary $\tilde{r}_{HMM}$.



Rysunek 8. Ścieżki Viterbiego dla szeregów czasowych x_t (linia ciągła) oraz y_t (linia przerywana) (przykład 2).

Miara korelacji HMM nie jest uniwersalna i nie może zostać użyta w analizie dowolnych szeregów. Wśród podstawowych ograniczeń należy wymienić przede wszystkim wymaganą długość szeregów czasowych; musi być ona większa niż liczba parametrów ukrytego modelu Markowa.

W następnych rozdziałach przedstawiona zostanie analiza porównawcza wspomnianych miar. Porównanie zostanie wykonane z użyciem rzeczywistych danych jakościowych (ankietowych).

4. Charakterystyka danych

W badaniu wykorzystane zostały odpowiedzi na pytania pochodzące z testu koniunktury w przemyśle przetwórczym, realizowanego comiesięcznie przez Instytut Rozwoju Gospodarczego SGH. Dane pochodzą z okresu od marca 1997 do listopada 2017 roku. W skład ankiety wchodzą następujące pytania:

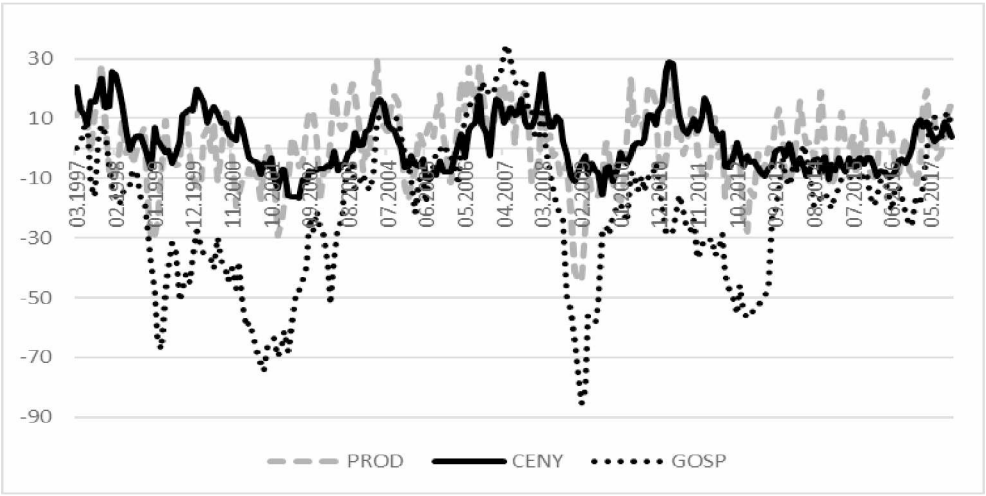
Pytanie 1 – wielkość produkcji (PROD),
Pytanie 2 – portfel zamówień ogółem (ZAM),
Pytanie 3 – portfel zamówień eksportowych (EKSP),
Pytanie 4 – poziom zapasów produkowanych wyrobów (ZAP),
Pytanie 5 – ceny produkowanych wyrobów (CENY),
Pytanie 6 – poziom zatrudnienia (ZAT),
Pytanie 7 – sytuacja finansowa przedsiębiorstwa (FIN),
Pytanie 8 – ogólna sytuacja gospodarki polskiej (GOSP).

Na każde z pytań możliwe są trzy odpowiedzi: wzrost, brak zmian lub spadek w porównaniu z poprzednim badaniem (miesiącem). Na podstawie odpowiedzi udzielonych przez respondentów wyznacza się saldo, tzn. różnicę między odsetkiem respondentów, którzy zgłosili wzrost wartości zmiennej objętej pytaniem, a odsetkiem respondentów, którzy zgłosili spadek. Salda te stanowią podstawę konstrukcji wskaźników koniunktury publikowanych przez IRG.

Wykres z szeregami sald odpowiedzi na pytania 1, 5 i 8 (PROD, CENY, GOSP) przedstawiony został na Rysunku 9. Rzuca się w oczy duża zmienność i różnorodność między nimi. Wszystkie osiem szeregów zostało poddanych badaniu, którego wyniki przedstawiono w następnym rozdziale.

5. Wyniki badania

Celem badania było zastosowanie miary korelacji HMM do szeregów sald testu koniunktury w przemyśle przetwórczym oraz porównanie wyników z wartościami współczynnika korelacji Pearsona. Biorąc pod uwagę liczbę pytań w ankiecie (osiem), liczba możliwych par różnych szeregów wynosi 28. Wartości współczynnika korelacji Pearsona wraz z wartościami p testów ich istotności zostały zebrane w Tabeli 1, a wartości współczynnika korelacji HMM w Tabeli 2. Dla celów porównawczych, w Tabeli 3, podano również wartości miary $\tilde{r}_{HMM}$. Poza nielicznymi wyjątkami są one większe od r_{HMM}.



Rysunek 9. Szereg czasowy sald odpowiedzi respondentów na pytania PROD (kreskowany), CENY (ciągły) oraz GOSP (kropkowany) od marca 1997 do listopada 2017.

Tabela 1. Współczynniki korelacji Pearsona wraz z wartościami *p* testu istotności.

r	ZAM	EKSP	ZAP	CENY	ZAT	FIN	GOSP
PROD	0,93 (5,8E-111)	0,82 (1,13E-61)	-0,27 (1,65E-05)	0,37 (2,44E-09)	0,56 (3,74E-22)	0,75 (1,07E-45)	0,62 (1,52E-27)
ZAM		0,88 (2,46E-80)	-0,39 (1,93E-10)	0,43 (1,3E-12)	0,67 (9,92E-34)	0,87 (8,07E-79)	0,76 (3,5E-48)
EKSP			-0,25 (8,02E-05)	0,45 (4,08E-14)	0,51 (4,12E-18)	0,74 (1,94E-44)	0,67 (6,59E-34)
ZAP				-0,06 (0,31)	-0,34 (2,51E-08)	-0,50 (2,84E-17)	0,31 (4,06E-07)
CENY					0,12 (0,0584)	0,39 (2,39E-10)	0,38 (9,6E-10)
ZAT						0,72 (1,1E-41)	0,73 (2,84E-43)
FIN							0,85 (3,17E-72)

Źródło: obliczenia własne na podstawie danych IRG SGH.

Tabela 2. Współczynniki korelacji r_{HMM}.

r_{HMM}	ZAM	EKSP	ZAP	CENY	ZAT	FIN	GOSP
PROD	0,66	0,41	0,51	0,51	0,65	0,77	0,57
ZAM		0,44	0,17	0,47	0,65	0,62	0,59
EKSP			0,49	0,56	0,64	0,67	0,62
ZAP				0,69	0,31	0,24	0,35
CENY					0,58	0,59	0,52
ZAT						0,49	0,52
FIN							0,46

Źródło: obliczenia własne na podstawie danych IRG SGH.

Tabela 3. Współczynniki alternatywnej korelacji $\tilde{r}_{HMM}$.

$\tilde{r}_{HMM}$	ZAM	EKSP	ZAP	CENY	ZAT	FIN	GOSP
PROD	0,68	0,62	0,45	0,65	0,61	0,72	0,67
ZAM		0,92	0,31	0,61	0,78	0,90	0,78
EKSP			0,33	0,63	0,71	0,83	0,78
ZAP				0,53	0,30	0,30	0,36
CENY					0,53	0,59	0,59
ZAT						0,87	0,79
FIN							0,88

Źródło: obliczenia własne.

Na poziomie istotności 0,05 wszystkie wartości współczynnika korelacji Pearsona okazały się istotne poza dwoma parami szeregów: (ZAP, CENY) oraz (CENY, ZAT). Wartości współczynnika r_{HMM} są nieco mniej zróżnicowane; najniższa wartość wynosi 0,17 dla pary (ZAM, ZAP), a największa to 0,77 dla pary (PROD, FIN). Z kolei wartości współczynnika korelacji Pearsona, które okazały się istotne, mieszczą się w zakresie – co do wartości bezwzględnej – od 0,25 dla pary (EKSP, ZAP) do 0,93 dla pary (PROD, ZAM). W wielu przypadkach wartości współczynnika korelacji HMM okazały się mniejsze od wartości współczynnika korelacji Pearsona. Precyzyjniejszych wniosków dostarczy analiza par szeregów (w Załączniku):
(PROD, ZAM) – Rysunek 10,
(PROD, CENY) – Rysunek 11,
(PROD, ZAT) – Rysunki 12 i 13,
(PROD, FIN) – Rysunek 14,
(ZAM, EKSP) – Rysunki 15 i 16,
(EKSP, GOSP) – Rysunek 17,

(ZAP, CENY) – Rysunek 18,
(CENY, ZAT) – Rysunek 19,
(FIN, GOSP) – Rysunek 20,
które zostały wybrane z uwagi na dostatecznie wysokie bądź niskie wartości któregokolwiek ze współczynników korelacji, albo też ze względu na przesłanki o charakterze ekonomicznym. Krótkie ich omówienie znajduje się w dalszej części rozdziału. Każdy z rysunków składa się z trzech wykresów. Pierwszy przedstawia analizowane szeregi, drugi szeregi po normalizacji, a trzeci szereg z_t (z kroku 2) oraz odpowiadającą mu ścieżkę Viterbiego.

Na Rysunku 10 przedstawione zostały szeregi sald odpowiedzi na pytania o wielkości produkcji i zamówień ogółem. Dla tej pary szeregów czasowych wartość współczynnika korelacji Pearsona jest najwyższa i wynosi 0,93. Oznacza to niemal doskonałe, dodatnie skorelowanie. Tymczasem stany 1 na ścieżce Viterbiego wyraźnie wskazują okresy, w których charakterystyki obu szeregów różniły się. Stąd wartość współczynnika korelacji HMM równa 0,66, choć nadal wysoka, jest znacznie niższa od wartości współczynnika korelacji Pearsona. Oceniając wzrokowo podobieństwo szeregów, wydaje się, że miara oparta na HMM oddaje podobieństwo bliższe rzeczywistości niż miara oparta na wzorze Pearsona.

Z kolei z oglądu Rysunku 12 wynika, że szeregi sald odpowiedzi na pytania o wielkości produkcji i zatrudnienia są bardzo podobne do siebie. Oba współczynniki korelacji dają zbliżone wyniki oceny stopnia podobieństwa ($r_{HMM} = 0{,}65$, $r = 0{,}56$). Stany na ścieżce Viterbiego identyfikują okresy podobieństwa oraz większego zróżnicowania szeregów. Wyraźnie rozdzielone są one cezurą 05.2004. Przed tą datą, według wskazań miary korelacji HMM, szeregi są znacznie mniej podobne niż później. Na Rysunku 13 przedstawiona jest ta sama para szeregów, ale tylko w okresie od maja 2004 roku. Zwróćmy uwagę, jak zmieniły się wartości współczynników korelacji. Wartość współczynnika korelacji Pearsona wzrosła do poziomu $r = 0{,}74$, natomiast wartość współczynnika korelacji HMM znacznie zmalała, do wartości $r_{HMM} = 0{,}32$. Mimo że ścieżka Viterbiego na całej rozpiętości szeregu (Rysunek 12) zawiera wyłącznie stany 0, to na skróconej ścieżce (Rysunek 13) stany 0 stanowią zaledwie 32% jej długości. Tę zaskakującą na pierwszy rzut oka niespójność nietrudno wytłumaczyć, a mianowicie stany na ścieżce Viterbiego dobierane są tak, aby zapewnić najbardziej prawdopodobną realizację łańcucha Markowa w całym okresie objętym badaniem. Zmiana tego okresu zmienia potencjalnie układ stanów. Innymi słowy, stan w konkretnym punkcie czasu dobierany jest optymalnie względem całego rozpatrywanego okresu. W pierwszym zatem przypadku

(Rysunek 12) po maju 2004 roku szeregi były podobne, biorąc pod uwagę okres od marca 1997 do listopada 2017 roku, w drugim zaś (Rysunek 13) szeregi zostały uznane za podobne w okresie od sierpnia 2013 roku, ale z perspektywy rozciągającej się od maja 2004 r. (a nie marca 1997 r.).

W przypadku szeregów sald odpowiedzi na pytania o wielkość produkcji i sytuację finansową przedsiębiorstwa (Rysunek 14) obie miary wskazują bardzo podobny i dość wysoki stopień korelacji ($r_{HMM} = 0{,}77, r = 0{,}75$), co wydaje się znajdować rozsądne uzasadnienie ekonomiczne. Podobnie jest zresztą w przypadku szeregów sald odpowiedzi na pytania o wielkość zamówień eksportowych i ogólny stan gospodarki polskiej (Rysunek 17) ($r_{HMM} = 0{,}62, r = 0{,}67$).

Z kolei Rysunek 15 przedstawia przykład szeregów (sald odpowiedzi na pytania o wielkości zamówień ogółem i eksportowych), dla których wartość współczynnika korelacji HMM jest dwukrotnie niższa od wartości współczynnika korelacji Pearsona ($r_{HMM} = 0{,}44,\ r = 0{,}88$). Zarazem wartość alternatywnej miary $\tilde{r}_{HMM}$ jest większa od wartości obu współczynników ($\tilde{r}_{HMM} = 0{,}92$). Zamówienia eksportowe stanowią część zamówień ogółem. Korelacja HMM jest w takim przypadku bardziej wrażliwa na zmiany w zmienności składu portfela zamówień. Zmienność ta od kwietnia 2006 roku okazała się, według miary HMM, inna dla zamówień eksportowych i ogółu zamówień. Widać to wyraźnie na wykresie szeregu z_t. Różnice pomiędzy szeregami łatwiej dostrzec na (powiększonym) wykresie szeregów po normalizacji (Rysunek 16). Wydaje się, że szeregi nie są tak podobne, jak na to wskazują miary r oraz $\tilde{r}_{HMM}$. Być może jednak, proponowana miara r_{HMM} niedoszacowuje stopień podobieństwa obu szeregów. Należy przy tym zwrócić uwagę, iż wysokie wartości współczynnika korelacji HMM występują znacznie rzadziej niż pozostałych dwóch współczynników.

Różnice między wartościami współczynników korelacji między pozostałymi ww. szeregami sald są bardzo duże. Wartość współczynnika korelacji Pearsona ($r = -0{,}06$) między szeregami sald odpowiedzi na pytania o wielkość zapasów wyrobów gotowych i ich cen (Rysunek 18) okazała się nieistotna statystycznie (p=0,31), a wartość współczynnika korelacji HMM jest zaskakująco wysoka ($r_{HMM} = 0{,}69$). Podobna jest różnica między wartościami współczynników korelacji: HMM ($r_{HMM} = 0{,}58$) i Pearsona ($r = 0{,}12$, statystycznie nieistotna – wartość p równa 0,0584) między szeregami sald odpowiedzi na pytania o ceny i wielkość zatrudnienia (Rysunek 19). Wydaje się, że w przypadku szeregów o dużej

zmienności (mierzoną amplitudą szeregu z_t) obie miary nie są odpowiednie. Duża jest również różnica między wartościami współczynników korelacji ($r_{HMM} = 0{,}46,\ r = 0{,}85$) między szeregami sald odpowiedzi na pytania o sytuację finansową przedsiębiorstwa i ogólną sytuację gospodarki polskiej (Rysunek 20). I w tym przypadku zmienność szeregu z_t jest wysoka. Powoduje to częste zmiany stanów na ścieżce Viterbiego, a w rezultacie dużą wrażliwość korelacji HMM (zmiany siły korelacji w czasie). Należy jednak zwrócić uwagę, że wartość $r = 0{,}85$ wydaje się zbyt wysoka, oceniając przebieg oryginalnych (czy też znormalizowanych) szeregów sald.

6. Wnioski

Przeprowadzona analiza prowadzi do następujących wniosków:

Przydatność proponowanej miary korelacji opartej na ukrytych modelach Markowa i ścieżkach Viterbiego poparta jest teoretycznymi właściwościami oraz empirycznymi przykładami.

Miara korelacji HMM nie jest uniwersalna, ale jej zakres stosowalności jest większy niż zakres stosowalności współczynnika korelacji liniowej Pearsona.

Miara korelacji HMM jest wrażliwa na wahania, zwłaszcza silne, wartości obserwowanej zmiennej w czasie, gdyż identyfikacja stanów na ścieżce Viterbiego jest uwarunkowana rozpiętością wartości zmiennej w próbie, a nie jej zmiennością lokalną.

Zaletą stosowania współczynnika korelacji HMM jest możliwość wskazania okresów podobieństwa i różnic między szeregami, a nadto wygodna interpretacja ekonomiczna.

Szeregi sald odpowiedzi na pytania testu koniunktury w przemyśle przetwórczym IRG SGH znacznie różnią się od siebie, biorąc pod uwagę wartości współczynnika korelacji HMM. Współczynnik korelacji Pearsona dla niektórych par szeregów wskazuje na bardzo wysoki stopień ich podobieństwa, co może przemawiać za rezygnacją z niektórych pytań ankiety.

Literatura

Bernardelli, M. (2014). Parallel deterministic procedure based on hidden Markov models for the analysis of economic cycles in Poland. *Roczniki Kolegium Analiz Ekonomicznych SGH*, *34*: 75-87.

Bernardelli, M. Hidden Markov models as a tool for assessing dependence of phenomena of an economic nature. *Acta Universitatis Lodziensis. Folia Oeconomica*. To be published.

Bernardelli, M., Dędys, M. (2014). The Viterbi path of hidden Markov models in an analysis of business tendency surveys, *Prace i Materiały Instytutu Rozwoju Gospodarczego SGH*.

Bernardelli M., Dędys M. (2017). Mapping the respondents' assessments in the RIED manufacturing tendency survey using the Viterbi paths. *Prace i Materiały Instytutu Rozwoju Gospodarczego*, *101*: 27-44.

Cappé O., Moulines E., Rydén T. (2005). *Inference in Hidden Markov Models*. Springer.

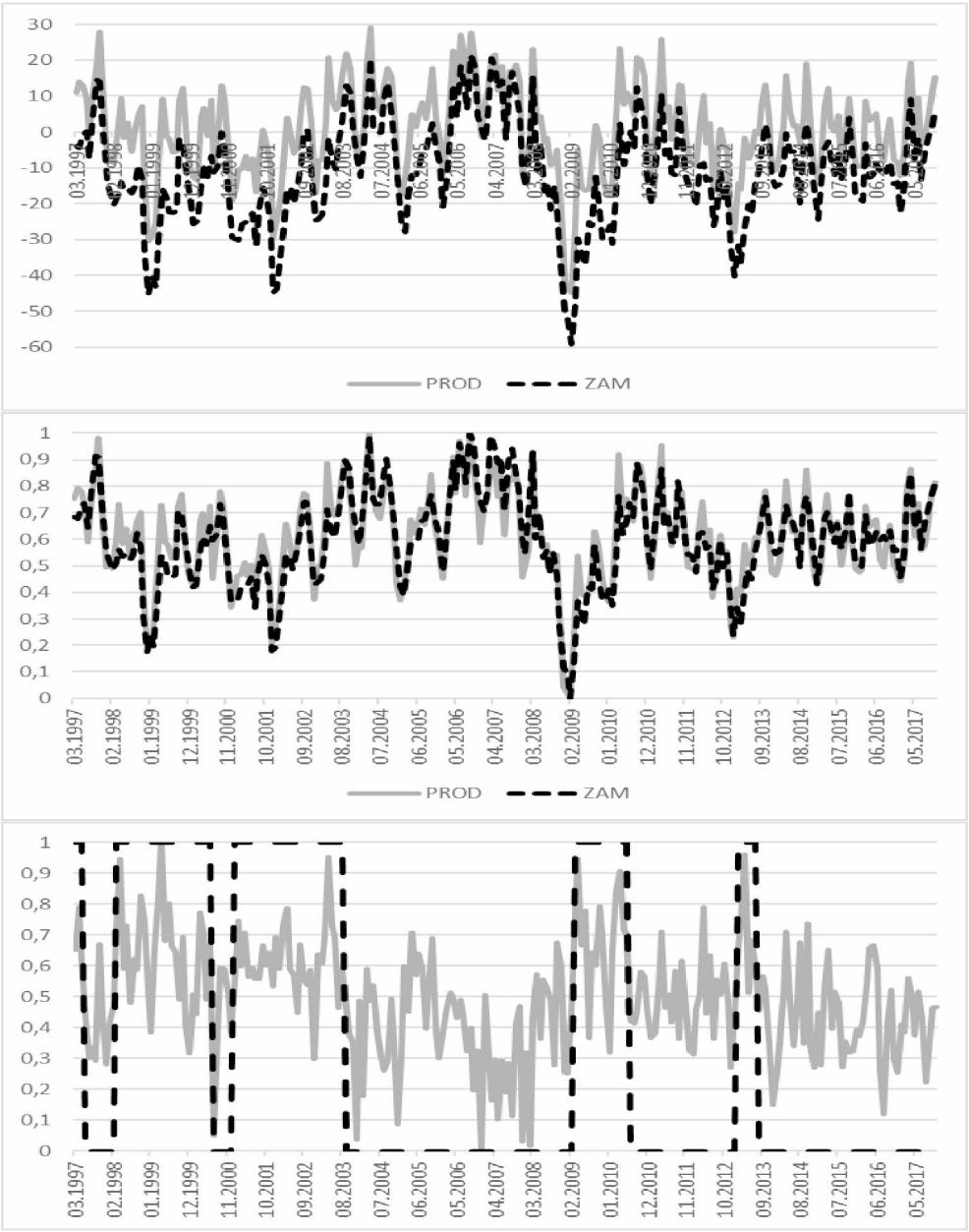
Kendall, M. G., Stuart, A. (1973), *The Advanced Theory of Statistics*, Volume 2: Inference and Relationship, Griffin.

Soper, H. E., Young, A. W., Cave, B. M., Lee, A., Pearson, K. (1917). On the distribution of the correlation coefficient in small samples. *Biometrika*, *11*: 328–413.

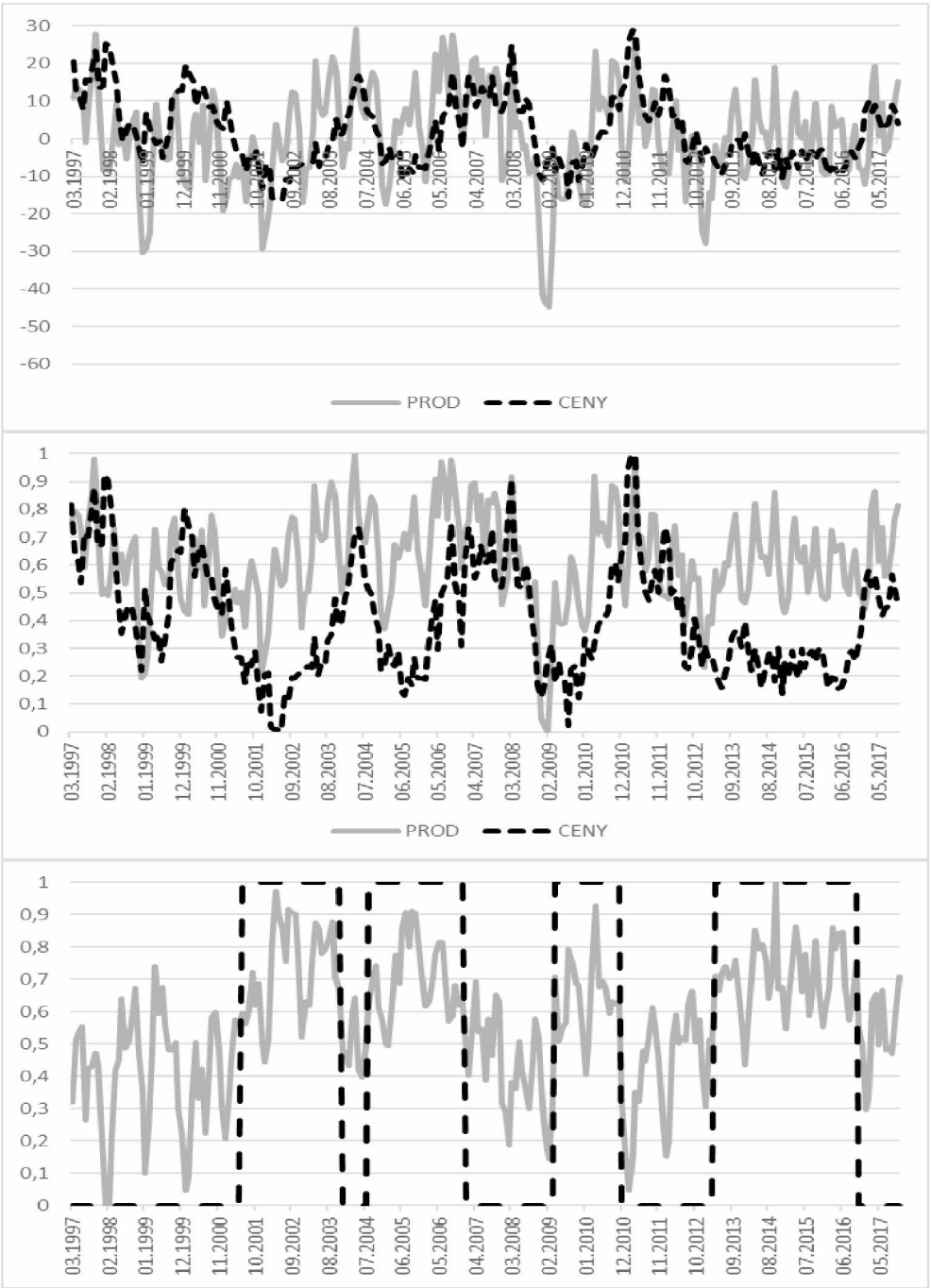
Székely, G. J., Rizzo M. L., Bakirov N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.*, *35*(6): 2769-2794.

Tjostheim D., Hufthammer K. O. (2013). Local Gaussian correlation: A new measure of dependence. *Journal of Econometrics*, *172*(1): 33-48.

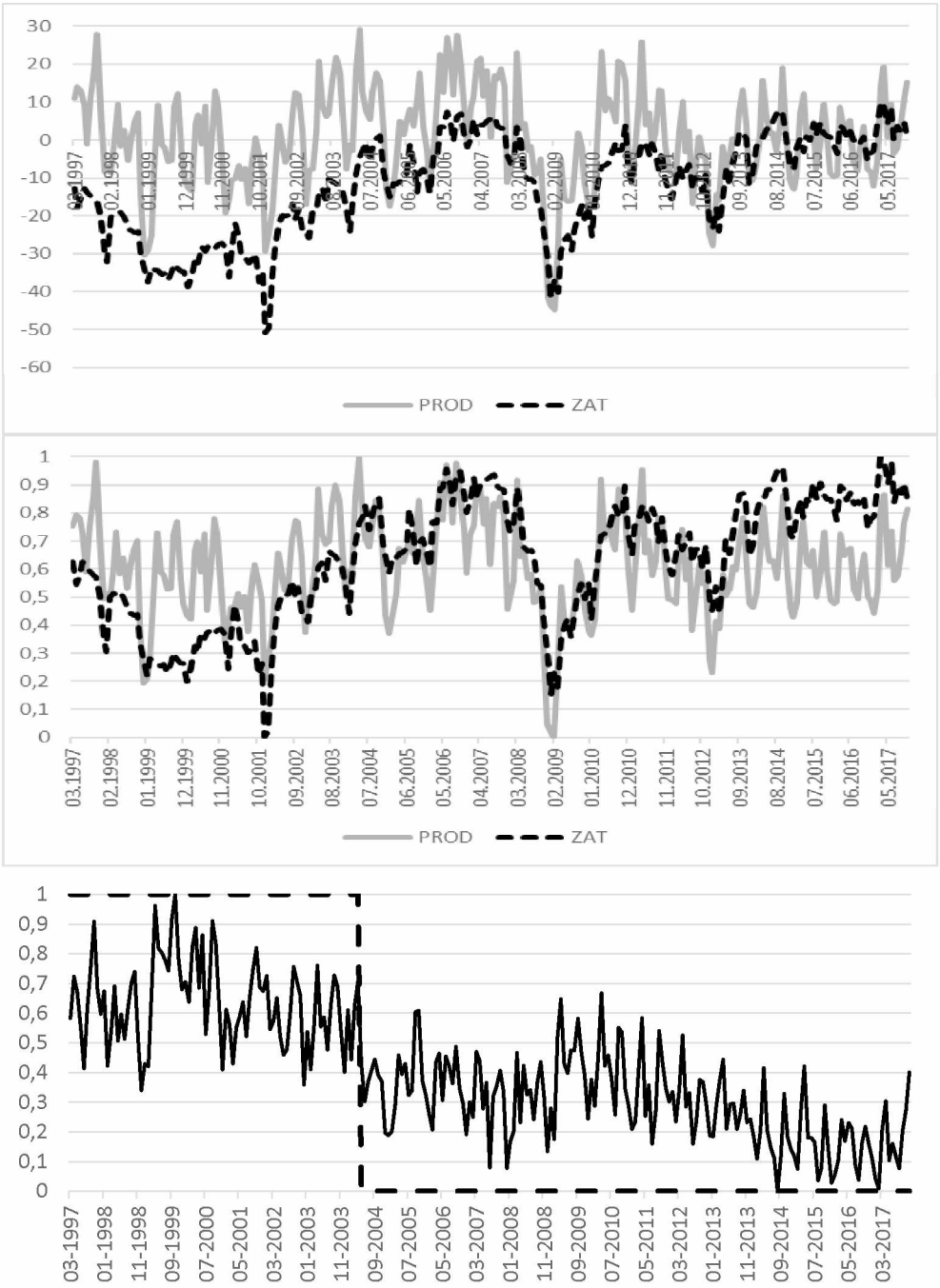
Załącznik



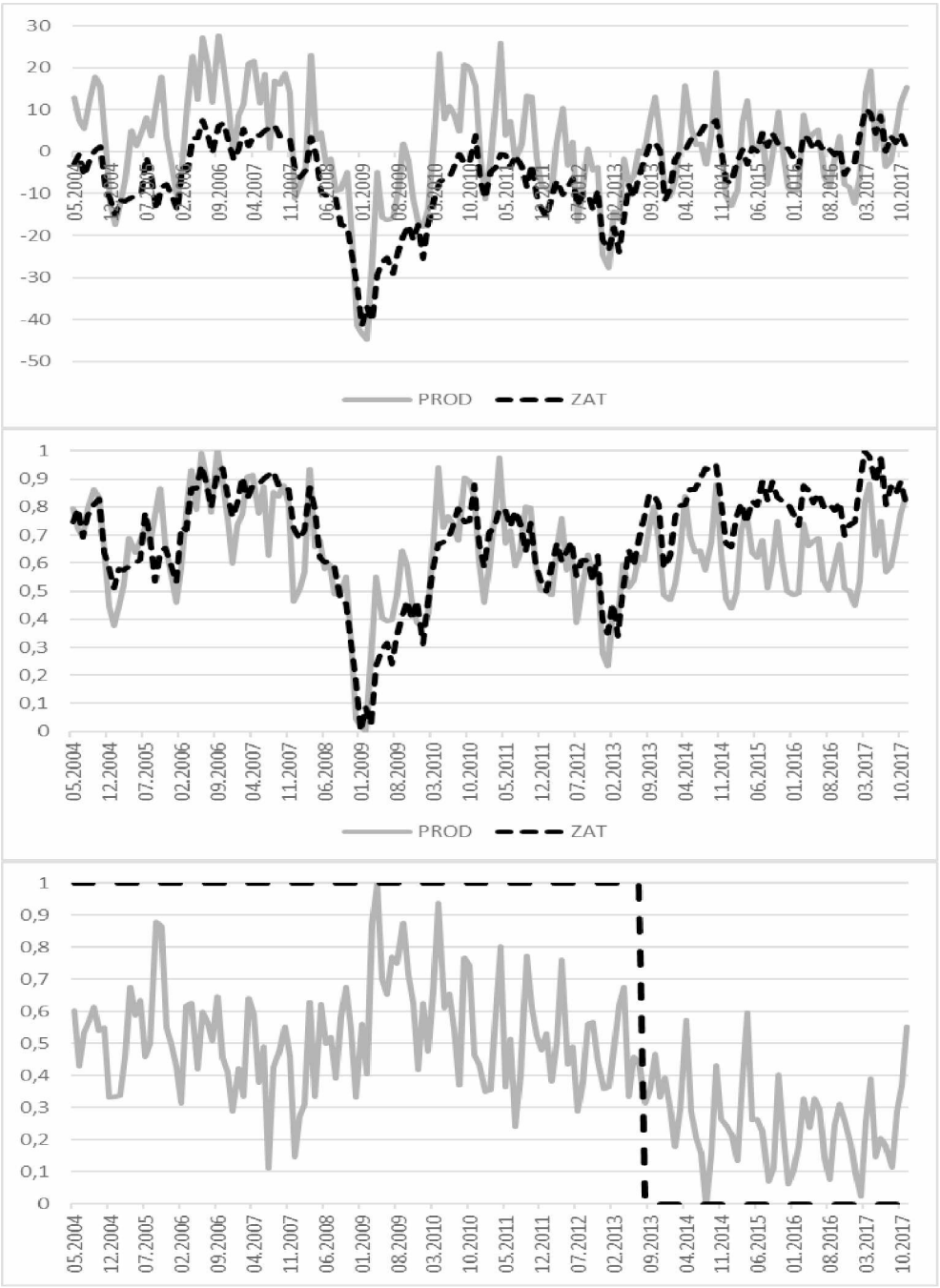
Rysunek 10. Wartości współczynnika korelacji HMM dla szeregów sald odpowiedzi na pytania o wielkości produkcji (PROD) i zamówień ogółem (ZAM), $r_{HMM} = 0{,}66$, $r = 0{,}93$ (p=5,82E-111), $\tilde{r}_{HMM} = 0{,}68$.



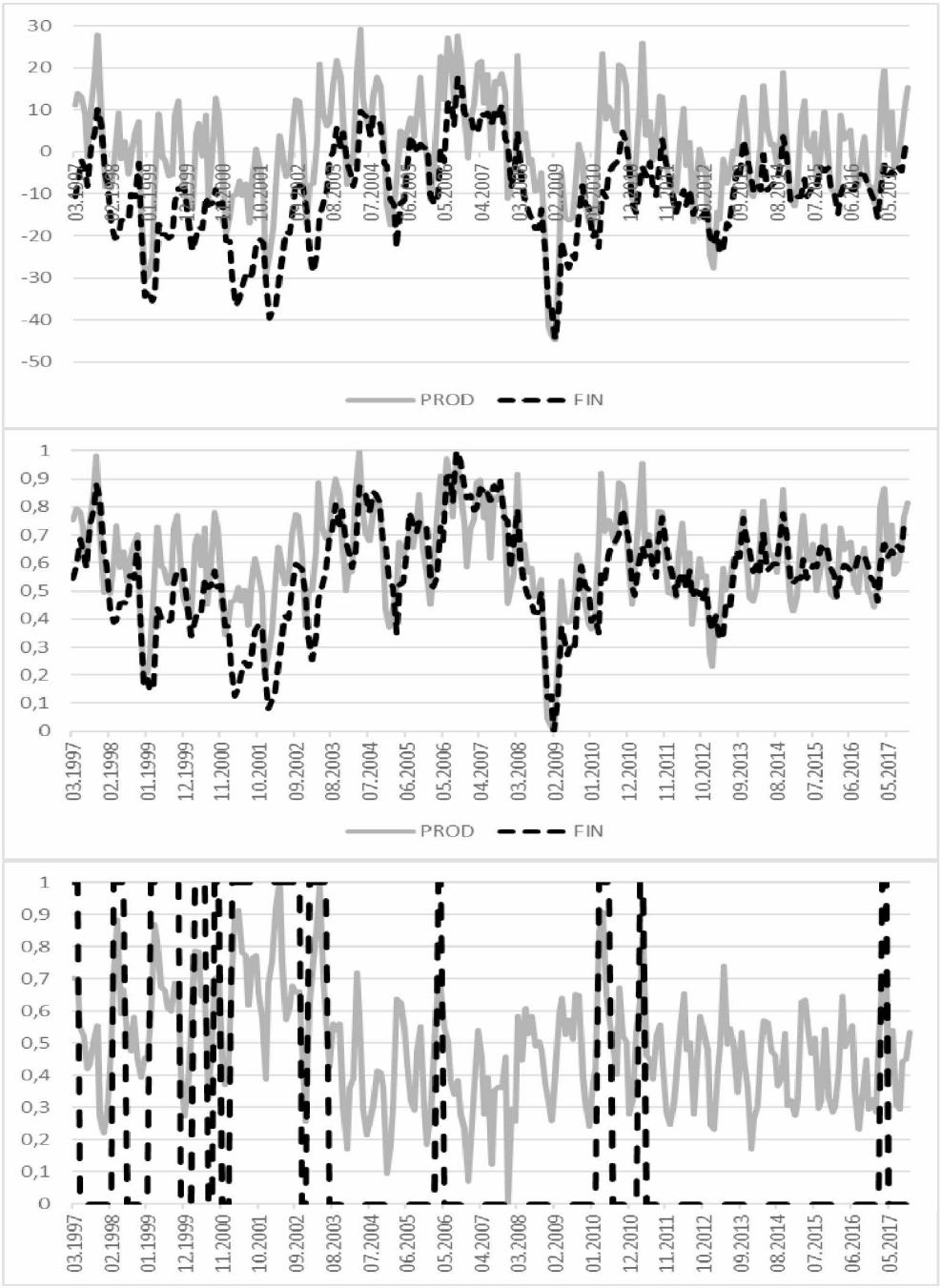
Rysunek 11. Wartości współczynnika korelacji HMM dla szeregów sald odpowiedzi na pytania o wielkość produkcji (PROD) i ceny wyrobów (CENY), $r_{HMM} = 0{,}51$, $r = 0{,}37$ (p=2,44E-09), $\tilde{r}_{HMM} = 0{,}65$.



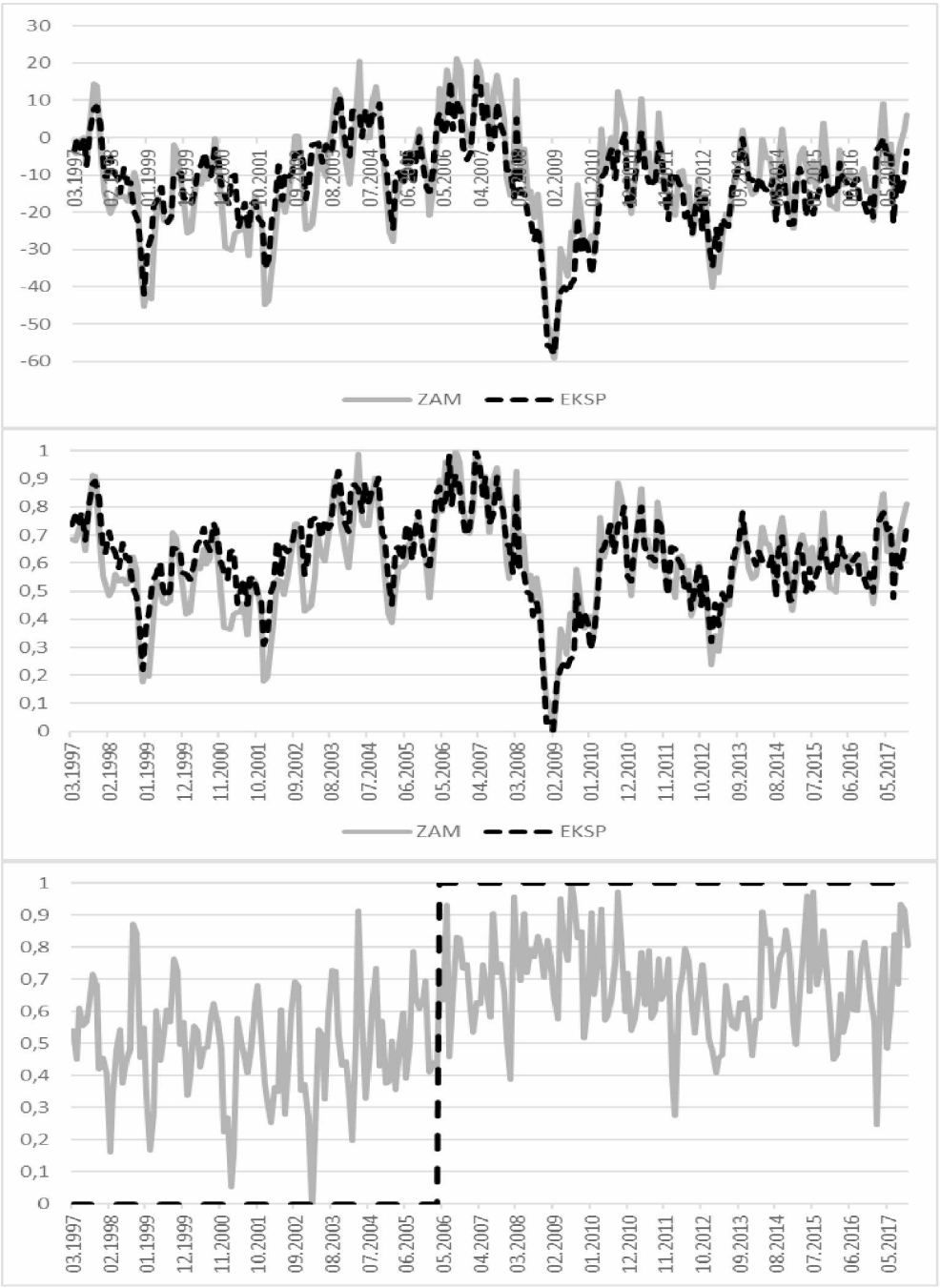
Rysunek 12. Wartości współczynnika korelacji HMM dla szeregów sald odpowiedzi na pytania o wielkości produkcji (PROD) i zatrudnienia (ZAT), $r_{HMM} = 0{,}65$, $r = 0{,}56$ (p=3,74E-22), $\tilde{r}_{HMM} = 0{,}61$.



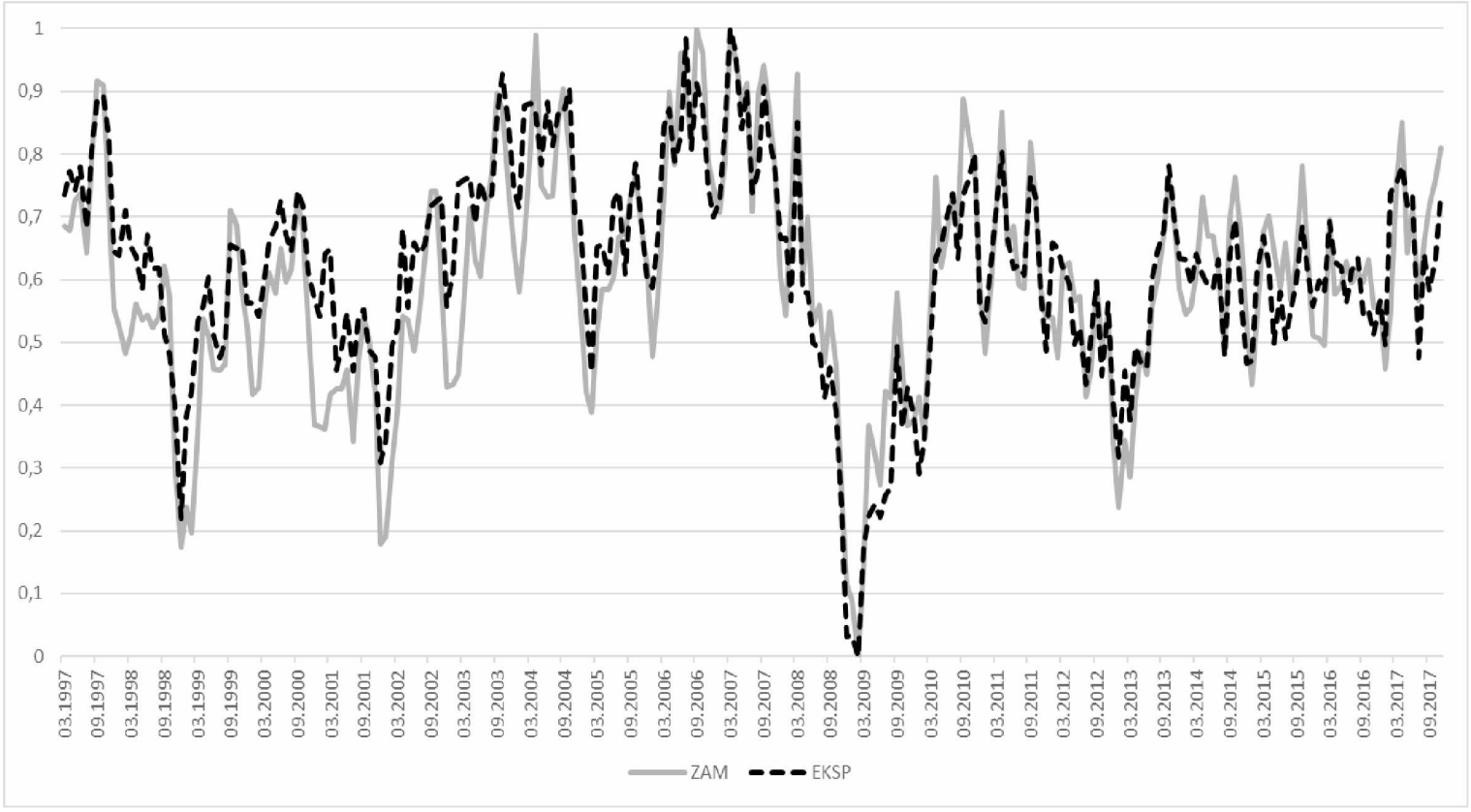
Rysunek 13. Wartości współczynnika korelacji HMM dla szeregów sald odpowiedzi na pytania o wielkości produkcji (PROD) i zatrudnienia (ZAT) z miesięcy od maja 2004 do listopada 2017 r., $r_{HMM} = 0{,}32$, $r = 0{,}74$ (p=2,44E-29), $\tilde{r}_{HMM} = 0{,}75$.



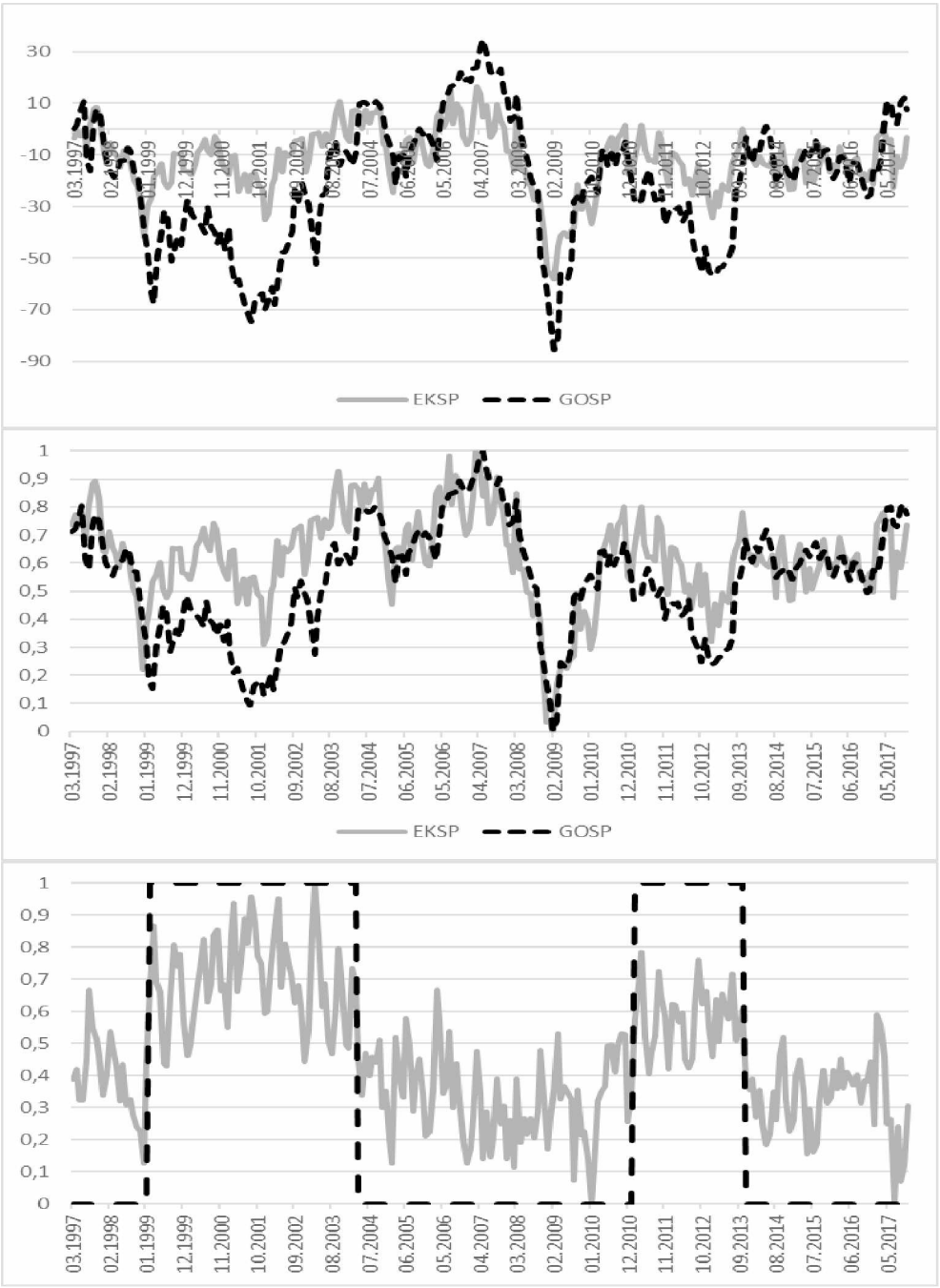
Rysunek 14. Wartości współczynnika korelacji HMM dla szeregów sald odpowiedzi na pytania o wielkość produkcji (PROD) i sytuację finansową przedsiębiorstwa (FIN), $r_{HMM} = 0{,}77, r = 0{,}75$ (p=1,07E-45), $\tilde{r}_{HMM} = 0{,}72$.



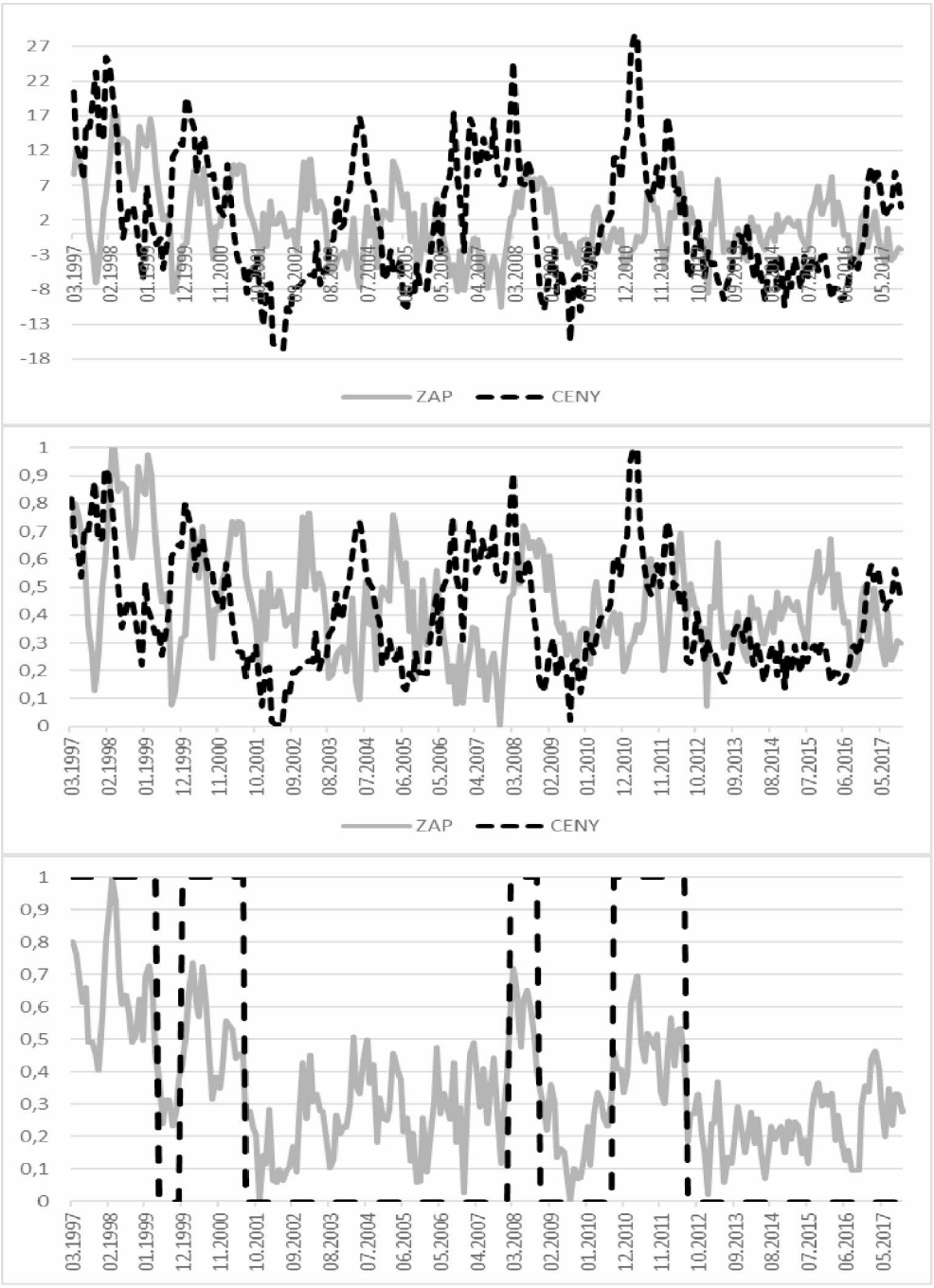
Rysunek 15. Wartości współczynnika korelacji HMM dla szeregów sald odpowiedzi na pytania o wielkości zamówień ogółem (ZAM) i eksportowych (EKSP), $r_{HMM} = 0{,}44$, $r = 0{,}88$ (p = 2,46E-80), $\tilde{r}_{HMM} = 0{,}92$.



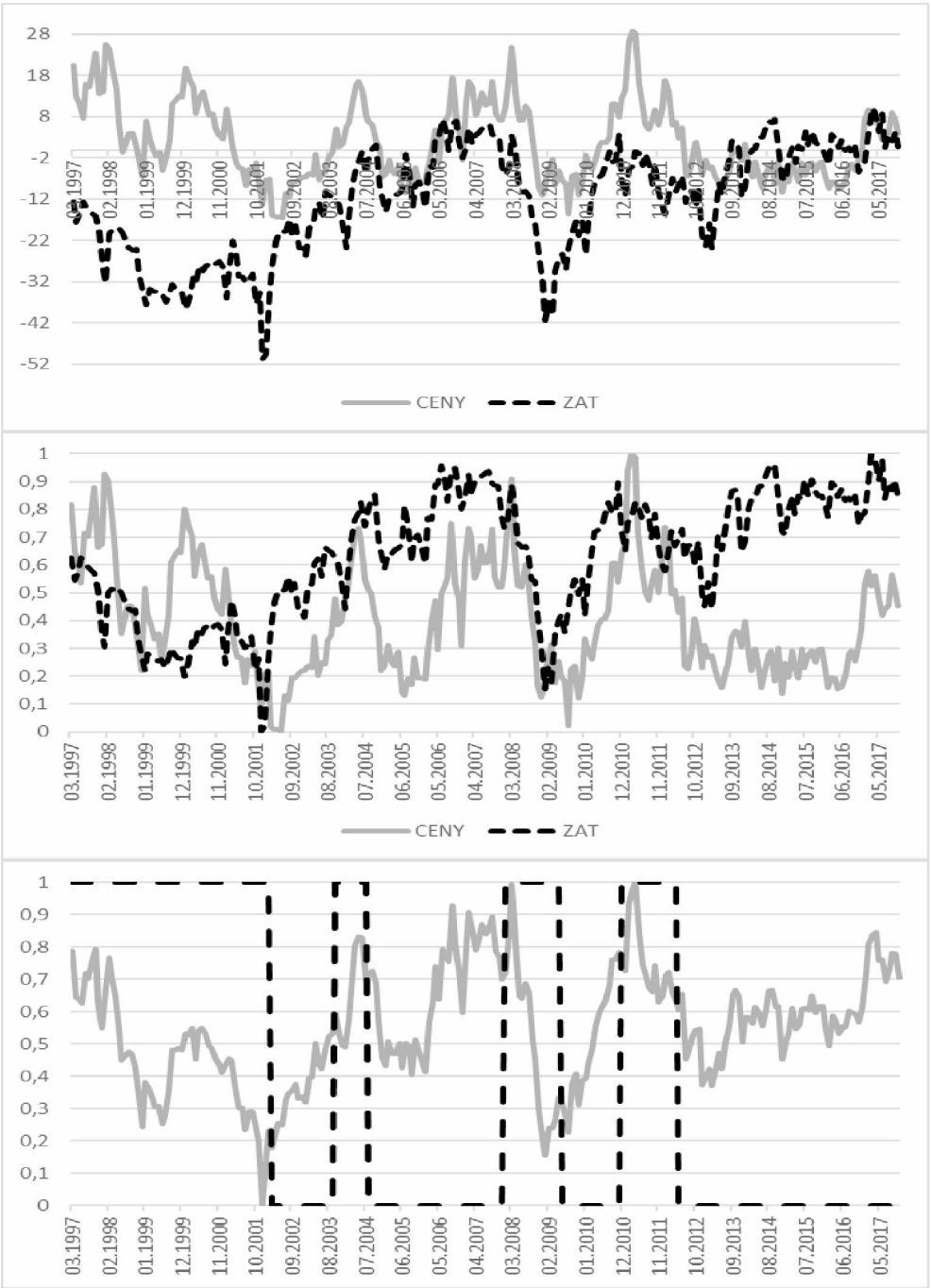
Rysunek 16. Szeregi sald odpowiedzi na pytania o wielkości zamówień ogółem (ZAM) i eksportowych (EKSP) po normalizacji, $r_{HMM} = 0{,}44$, $r = 0{,}88$ (p=2,46E-80), $\tilde{r}_{HMM} = 0{,}92$.



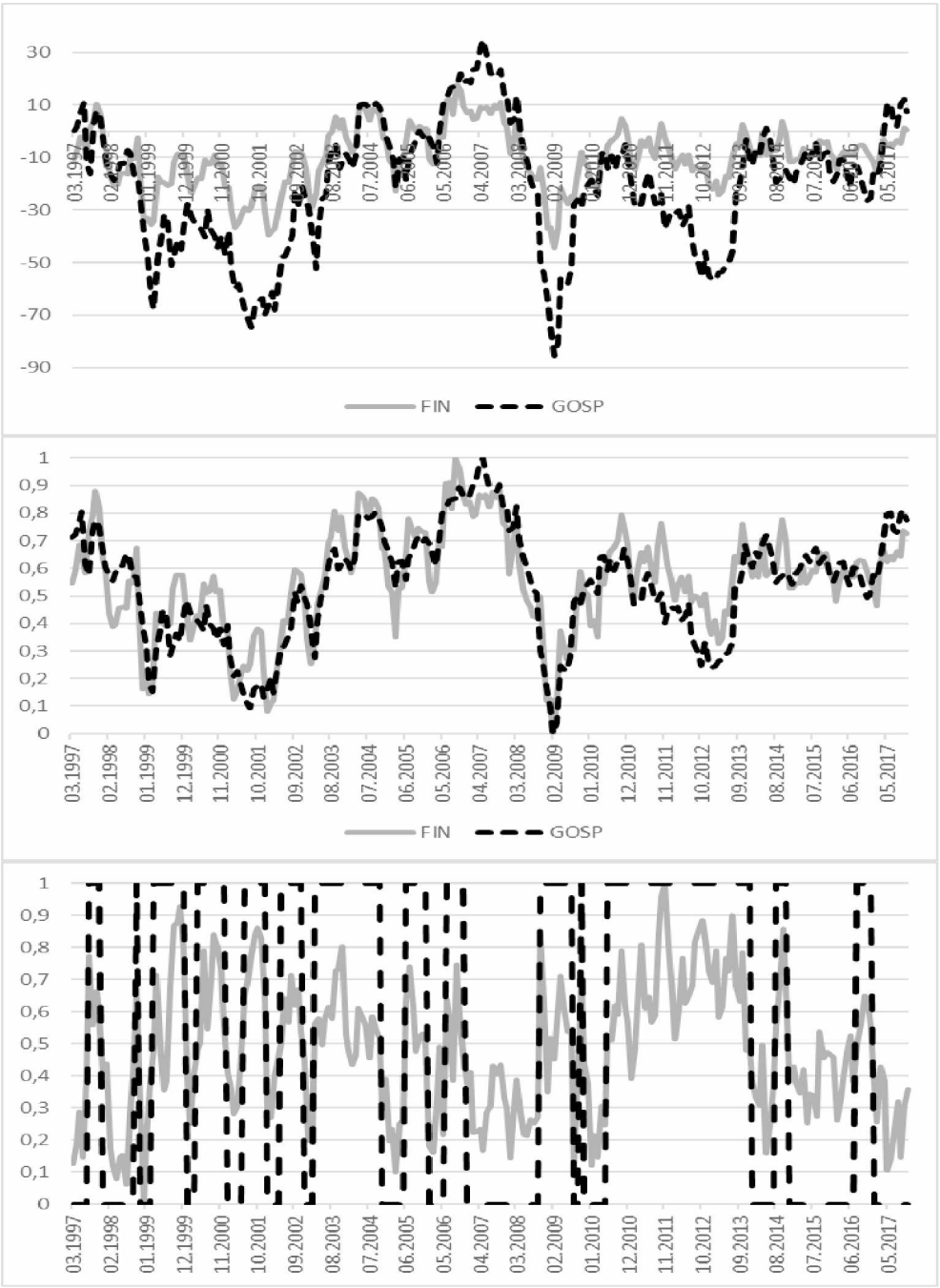
Rysunek 17. Wartości współczynnika korelacji HMM dla szeregów sald odpowiedzi na pytania o wielkość zamówień eksportowych (EKSP) i ogólną sytuację gospodarki polskiej (GOSP), $r_{HMM} = 0{,}62$, $r = 0{,}67$ (p=6,59E-34), $\tilde{r}_{HMM} = 0{,}78$.



Rysunek 18. Wartości współczynnika korelacji HMM dla szeregów sald odpowiedzi na pytania o wielkość zapasów (ZAP) i ceny wyrobów gotowych (CENY), $r_{HMM} = 0{,}69$, $r = -0{,}06$ (p=0,31), $\tilde{r}_{HMM} = 0{,}53$.



Rysunek 19. Wartości współczynnika korelacji HMM dla szeregów sald odpowiedzi na pytania o ceny wyrobów (CENY) i wielkość zatrudnienia (ZAT), $r_{HMM} = 0{,}58$, $r = 0{,}12$ (p=0,0584), $\tilde{r}_{HMM} = 0{,}53$.



Rysunek 20. Wartości współczynnika korelacji HMM dla szeregów sald odpowiedzi na pytania o sytuację finansową przedsiębiorstwa (FIN) i ogólną sytuację gospodarki polskiej (GOSP), $r_{HMM} = 0{,}46$, $r = 0{,}85$ (p=3,17E-72), $\tilde{r}_{HMM} = 0{,}88$.